BISS Biodiversity Information Science and Standards

OPEN ACCESS

Conference Abstract

The Future of Natural History Transcription: Navigating AI advancements with VoucherVision and the Specimen Label Transcription Project (SLTP)

William N Weaver[‡], Kyle Lough[‡], Stephen A Smith[‡], Brad Ruhfel[‡]

‡ University of Michigan, Ann Arbor, United States of America

Abstract

Natural history collections are critical reservoirs of biodiversity information but collections staff are constantly grappling with substantial backlogs and limited resources. The task of transcribing specimen label text into searchable databases requires a significant amount of time, manual labor, and funding. To address this challenge, we introduce VoucherVision, a tool harnessing the capabilities of several Large Language Models (LLMs; Naveed et al. 2023) to augment specimen label transcription. The VoucherVision tool automates laborious components of the transcription process, leveraging an Optical Character Recognition (OCR) system and LLMs to convert unstructured label text into appropriate data formats compatible with database ingestion. VoucherVision uses a combination of structured output parsers and recursive re-prompting strategies to ensure consistency and quality of the LLM-formatted text, significantly reducing errors.

Integration of VoucherVision with the University of Michigan Herbarium's transcription workflow resulted in a significant reduction in per-image transcription time, suggesting significant potential advantages for collections workflows. VoucherVision offers promising strides towards efficient digitization, with curatorial staff playing critical roles in data quality

assurance and process oversight. Emphasizing the importance of knowledge sharing, the University of Michigan Herbarium is backing the Specimen Label Transcription Project (SLTP), which will provide open access to benchmarking datasets, fine-tuned models, and validation tools to rank the performance of different methodologies, LLMs, and prompting strategies. In the rapidly evolving landscape of Artificial Intelligence (AI) development, we recognize the profound potential of diverse contributions and innovative methodologies to redefine and advance the transformation of curatorial practices, catalyzing an era of accelerated digitization in natural history collections.

An early, public version of VoucherVision is available to try here: https://voucher vision.azurewebsites.net/

Keywords

large language models, herbarium, specimen digitization, natural language processing

Presenting author

William Weaver

Presented at

TDWG 2023

Acknowledgements

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Barnes N, Mian A (2023) A Comprehensive Overview of Large Language Models. arXiv. arXiv:2307.06435 [cs]. URL: http://arxiv.org/abs/2307.06435